Solve the following problems (taken from your textbook).

2.16 Table 2.12 comes from one of the first studies of the link between lung cancer and smoking, by Richard Doll and A. Bradford Hill. In 20 hospitals in London, UK, patients admitted with lung cancer in the previous year were queried about their smoking behavior. For each patient admitted, researchers studied the smoking behavior of a noncancer control patient at the same hospital of the same sex and within the same 5-year grouping on age. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

Table 2.12. Data for Problem 2.16

	Lung Cancer	
Have Smoked	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Based on data reported in Table IV, R. Doll and A. B. Hill, *Br. Med. J.*, 739–748, September 30, 1950.

a. Identify the response variable and the explanatory variable.

b. Identify the type of study this was.

c. Can you use these data to compare smokers with nonsmokers in terms of the proportion who suffered lung cancer? Why or why not?

d. Summarize the association, and explain how to interpret it.

2.18 Table 2.13 shows data from the 2002 General Social Survey cross classifying a person's perceived happiness with their family income. The table displays the observed and expected cell counts and the standardized residuals for testing independence.

a. Show how to obtain the estimated expected cell count of 35.8 for the first cell.

b. For testing independence, $X^2 = 73.4$. Report the df value and the P-value, and interpret.

c. Interpret the standardized residuals in the corner cells having counts 21 and 83.

d. Interpret the standardized residuals in the corner cells having counts 110 and 94.

Table 2.13. Data for Problem 2.18, with Estimated Expected Frequencies and Standardized Residuals

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above average	21 35.8 −2.973	159 166.1 −0.947	110 88.1 3.144
Average	53 79.7 −4.403	372 370.0 0.224	221 196.4 2.907
Below average	94 52.5 7.368	249 244.0 0.595	83 129.5 −5.907

3.7 Access the horseshoe crab data in Table 3.2 at www.stat.ufl.edu/~aa/intro-cda/appendix.html. Let $Y = 1$ if a crab has at least one satellite, and let $Y = 0$ otherwise. Using weight as the predictor, fit the linear probability model.

a. Use ordinary least squares. Interpret the parameter estimates. Find the predicted probability at the highest observed weight of 5.20 kg. Comment.

b. Attempt to fit the model using ML, treating Y as binomial. What does your software report? [The failure is due to a fitted probability falling outside the (0, 1) range.]

c. Fit the logistic regression model. Show that the estimated logit at a weight of 5.20 kg equals 5.74. Show that $\hat{\pi} = 0.9968$ at that point by checking that $\log[\hat{\pi}/(1-\hat{\pi})] = 5.74$ when $\hat{\pi} = 0.9968$.

For the problem 3.7, write in R:

Crabs_dat_Table_203 <- read_table("http://users.stat.ufl.edu/~aa/intro-cda/data/Crabs.dat(Table%203.2)", col_types = cols(y = col_factor(levels = c("0", "1")))).

Alternatively, download the file in R format "Crabs_dat_Table_203.Rd" from your reserved space and use the load() function

3.8 Refer to the previous exercise for the horseshoe crab data.

a. Report the fit for the probit model, with weight predictor.

b. Find $\hat{\pi}$ at the highest observed weight, 5.20 kg.

c. Describe the weight effect by finding the difference between the $\hat{\pi}$ values at the upper and lower quartiles of weight, 2.85 and 2.00 kg.

d. Interpret the parameter estimates using characteristics of the normal *cdf* that describes the response curve.

3.16 One question in a recent General Social Survey asked subjects how many times they had had sexual intercourse in the previous month.

a. The sample means were 5.9 for males and 4.3 for females; the sample variances were 54.8 and 34.4. Does an ordinary Poisson GLM seem appropriate? Explain.

b. The GLM with log link and a dummy variable for gender (1 = males, 0 = females) has gender estimate 0.308. The SE is 0.038 assuming a Poisson distribution and 0.127 assuming a negative binomial model. Why are the SE values so different?

c. The Wald 95% confidence interval for the ratio of means is (1.26, 1.47) for the Poisson model and (1.06, 1.75) for the negative binomial model. Which interval do you think is more appropriate? Why?

6.6 Does marital happiness depend on family income? For the 2002 General Social Survey, counts in the happiness categories (not, pretty, very) were (6, 43, 75) for below average income, (6, 113, 178) for average income, and (6, 57, 117) for above average income. Table 6.15 shows output for a baseline-category logit model with very happy as the baseline category and scores {1, 2, 3} for the income categories.

a. Report the prediction equations from this table.

b. Interpret the income effect in the first equation.

c. Report the Wald test statistic and P-value for testing that marital happiness is independent of family income. Interpret.

d. Does the model fit adequately? Justify your answer.

e. Estimate the probability that a person with average family income reports a very happy marriage.

Table 6.15. Output on Modeling Happiness for Problem 6.6

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	3.1909	2	1.5954	0.2028
Pearson	3.1510	2	1.5755	0.2069

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.9439	2	0.6238
Wald	0.9432	2	0.6240

Analysis of Maximum Likelihood Estimates

Parameter	happy	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	−2.5551	0.7256	12.4009	0.0004
Intercept	2	1	−0.3513	0.2684	1.7133	0.1906
income	1	1	−0.2275	0.3412	0.4446	0.5049
income	2	1	−0.0962	0.1220	0.6210	0.4307

Use R (or Python) to make computations, estimate models and plot figures. I want two files: a big R script that allows me to reproduce your results, and a unique pdf file that contains your homework. I do not want just the solution. Please explain what you are doing and interpret the results.

Please send your teacher by mail the assignment by next December 24.

Good luck.